

Article Watch: December 2014

Clive A. Slaughter

GRU-UGA Medical Partnership, Athens GA

This column highlights recently published articles that are of interest to the readership of this publication. We encourage ABRF members to forward information on articles they feel are important and useful to Clive Slaughter, Georgia Regents University-University of Georgia Medical Partnership, 1425 Prince Ave., Athens, GA 30606, USA. Phone: 706-713-2216; Fax: 706-713-2221; E-mail: cslaught@uga.edu; or to any member of the Editorial Board. Article summaries reflect the reviewer's opinions and not necessarily those of the association.

DNA SEQUENCING AND CHARACTERIZATION

SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* 32;2014:903–914.

Li S, Labaj P P, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C, Thierry-Mieg D, Thierry-Mieg J, Kreil D P, Mason C E. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology* 32;2014:888–895.

Two papers describe the results of a multisite, cross-platform study of the performance of RNA-seq. The study has been conducted by the Sequencing Quality Control (SEQC) project, coordinated by the U.S. Food and Drug Administration. Two RNA samples were analyzed: the Universal Human Reference RNA and the Human Brain Reference RNA. They were spiked with synthetic RNA from the External RNA Control Consortium (ERCC). The samples were assessed individually and combined in known ratios. The platforms tested were Illumina HiSeq, Life Technologies SOLiD, and Roche 454. Differential expression profiling and junction discovery results were compared with analyses performed by microarray and quantitative PCR (qPCR). The consortium's publication confirms that RNA-seq enables discovery of verifiable splicing patterns. Relative expression patterns are ascertained accurately and reproducibly across sites and platforms, but absolute measurements are inaccurate, as indeed, are the results of microarrays. Gene-specific biases occur with all platforms, including qPCR. Li et al. identify GC content, gene coverage, sequencing error rate, and insert size to be sources of error that contribute to irrepro-

ducibility. Large numbers of supposedly differentially expressed genes are identified from the data at different sites, even though the self-same sample is being analyzed, indicating that library preparation is a major source of false-positives. This highlights the need for stringent normalization methods. These studies represent a benchmark for the performance of RNA-seq that will be of interest to all investigators involved in gene-expression analysis.

Wang C, Gong B, Bushel P R, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Labaj P P, Kreil D P, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian H-R, Lancashire L J, Bessarabova M, Nikolsky Y, Furlanello C, Chierici M, Albanese D, Jurman G, Riccadonna S, Filosi M, Visintainer R, Zhang K K, Li J, Hsieh J-H, Svoboda D L, Fuscoe J C, Deng Y, Shi L, Paules R S, Auerbach S S, Tong W. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology* 32;2014:926–932.

Wang et al. adopt a different study design, although also under the umbrella of the SEQC project. They study differential expression of genes in rat liver upon exposure to 27 chemicals with various different modes of action. Samples are tested with the Illumina RNA-seq and Affymetrix microarray platforms. The results show that the degree of concordance between the platforms depends on the degree of perturbation elicited by the treatment. RNA-seq is found to be better than microarrays for detecting weakly expressed genes. Finally, predictive models constructed from the data to classify the mode of action of the various chemicals are similar for the two platforms.

Li S, Tighe S W, Nicolet C M, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer P A, Gao Y, Kim D, Boland J, Hicks B, Kim R, Chhangawala S, Jafari N, Raghavachari

doi: 10.7171/jbt.14-2504-005

N, Gandara J, Garcia-Reyero N, Hendrickson C, Roberson D, Rosenfeld J, Smith T, Underwood J G, Wang M, Zumbo P, Baldwin D A, Grills G S, Mason C E. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology* 32;2014:915–925.

The Association of Biomolecular Resource Facilities reports a separate study of the performance of RNA-seq that complements the SEQC findings. It extends the number of instrument platforms, additionally including Illumina HiSeq 2500, Pacific Biosciences RS, and Life Technologies Personal Genome Machine (PGM) and Proton. The authors observe that deeper sampling is needed to reveal lower-abundance transcripts and splice junctions. Given that deep sampling is not presently cost-effective with long-read platforms, such as Pacific Biosciences and Roche 454, the best discovery platforms for low-abundance targets are the shorter-read platforms. However, longer reads help detection of splice junctions. The study also compares libraries prepared by various methods and indicates that poly-A enrichment, ribo-depletion, and even libraries made from severely degraded RNAs provide comparable results, given sufficient depth of coverage.

Risso D, Ngai J, Speed T P, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* 32;2014:896–902.

Risso et al. evaluate the performance of spike-in controls for RNA-seq. With the use of the ERCC spike-in controls, they conclude that the spike-ins are not reliable enough to be used in standard global-scaling or regression-based normalization procedures. They propose instead a normalization strategy that adjusts for variation from unknown sources, such as library preparation. The strategy involves factor analysis of counts for suitable sets of control genes (e.g., ERCC) or suitable samples (e.g., replicate libraries). The authors show that this alternative approach enhances the accuracy of expression fold-change measurements and the reliability of tests for differential expression. Investigators will wish to ensure that spike-in standards behave as expected in their experiments and to develop stable controls that are robust to differences in library composition and library-preparation methods.

PROTEINS—PURIFICATION AND CHARACTERIZATION

Nannenga B L, Shi D, Leslie A G, Gonen T. High-resolution structure determination by continuous-rotation data collection in MicroED. *Nature Methods* 11;2014:927–930.

The capability to use microcrystals for X-ray diffraction studies rather than the conventional millimeter-sized

crystals is expected to expand the range of proteins amenable to structural analysis significantly. Developments in X-ray free-electron lasers for this purpose have been encouraging, but many thousands of microcrystals are required by methods using them, and availability of the instrumentation is limited. Here, Nannenga et al. provide a progress report on developments in electron diffraction methods for high-resolution structure determination with microcrystals. Crystals of lysozyme, $\sim 2 \mu\text{m}$ in diameter and $0.1\text{--}0.6 \mu\text{m}$ in thickness, are loaded onto an electron microscope grid and diffracted at cryogenic temperatures. To improve accuracy, the microscope stage is rotated at a constant rate of $\sim 0.09^\circ\text{s}^{-1}$, and diffraction data from single microcrystals are recorded as a movie. Typically, 44° of data is collected before reaching the radiation dose limit for a given microcrystal. Visible reflections beyond 2 \AA are detected. Data collected using this methodology can be processed by standard programs developed for X-ray data. After phasing and refinement, the data enable the structure of lysozyme at 2.5 \AA resolution to be reported. These results encourage the expectation that microelectron diffraction will become a widely applicable method for structure analysis.

PROTEOMICS

de Graaf E L, Giansanti P, Altelaar A F, Heck A J. Single-step enrichment by Ti^{4+} -IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. *Molecular & Cellular Proteomics* 13;2014:2426–2434.

Matheron L, van den Toorn H, Heck A J, Mohammed S. Characterization of biases in phosphopeptide enrichment by Ti^{4+} -immobilized metal affinity chromatography and TiO_2 using a massive synthetic library and human cell digests. *Analytical Chemistry* 86;2014:8312–8320.

For mass spectral phosphoproteomic analysis, phosphopeptides are usually enriched from digests of cell proteins either by immobilized metal ion affinity chromatography (IMAC) or by titanium dioxide chromatography. Recently, these alternative approaches have been combined in a Ti^{4+} -IMAC enrichment scheme. Signal transduction networks may display a rapid, dynamic change in response to stimuli, so phosphoproteomic workflows must be streamlined to enable analysis of large numbers of samples to describe such dynamism adequately. Nevertheless, the complexity of signal transduction networks demands that the phosphorylation of very large numbers of protein target sites must be measured in each sample. In a study of phosphoproteome dynamics in Jurkat T cells stimulated by prostaglandin E_2 , de Graaf et al. combine Ti^{4+} -IMAC phosphopeptide enrichment with a workflow in which

single ultra-HPLC analysis of each sample is performed on-line with a Thermo Fisher Scientific LTQ Orbitrap Elite mass spectrometer to provide label-free quantification. They analyze >24,000 phosphorylation sites in total with this methodology. More than 10,000 of these sites are quantified in at least two out of three biological replicates. In a separate study by the same research group, Matheron et al. use a library of 23,000 synthetic phosphopeptides, present in roughly equal abundance, to characterize the biases introduced by the Ti^{4+} -IMAC phosphopeptide-enrichment methodology. They find the biases to be similar to those encountered in titanium dioxide chromatography. These studies illustrate the very large scale of experimentation required to track cellular responses to signaling stimuli.

Merrill A E, Hebert A S, MacGilvray M E, Rose C M, Bailey D J, Bradley J C, Wood W W, El Masri M, Westphall M S, Gasch A P, Coon J J. NeuCode labels for relative protein quantification. *Molecular & Cellular Proteomics* 13;2014: 2503–2512.

Merrill et al. demonstrate the capabilities of their research group's new multiplex approach to stable isotope labeling by amino acids in cell culture (SILAC). In this approach, neutron-encoded mass defects are exploited to make isotopologues of lysine, differing by just 36 mDa. These isotopologues are discriminated on a chromatographic time-scale by high-resolution mass spectrometry in an Orbitrap instrument. In the present paper, the authors describe methods for the synthesis of six such isotopologues and use them in studies of yeast cells exposed to salt stress. Four thousand proteins are quantified. The methodology is shown to extend the dynamic range of the SILAC approach, thereby enhancing its sensitivity and discovery potential. The authors further expand the plexing capacity to 18 with the addition of chemical labels.

FUNCTIONAL GENOMICS AND PROTEOMICS

Wu X, Scott D A, Kriz A J, Chiu A C, Hsu P D, Dadon D B, Cheng A W, Trevino A E, Konermann S, Chen S, Jaenisch R, Zhang F, Sharp P A. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology* 32;2014:670–676.

Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature Biotechnology* 32;2014:677–683.

Two groups here investigate the site specificity of gene editing by clustered regularly interspaced short palindromic repeat (CRISPR)-associated (Cas) proteins, guided by short-guide RNAs. This CRISPR-Cas methodology is

now used extensively for targeted genome editing, disease gene repair, and knock-in of specific tags. Both groups map the genome-wide binding of inactive Cas9, loaded with examples of short-guide RNA. Somewhat alarmingly, they discover that binding sites number between tens and thousands. These binding sites are enriched in open chromatin regions. However, off-target production of indels by active Cas9 occurs at much lower frequency than off-target binding, rarely rising above background levels. This favorable discrepancy between cleavage and binding is attributed to a requirement for more extensive base pairing to trigger cleavage. Nevertheless, investigators are advised to minimize potential off-target effects by cotargeting genes with several short-guide RNAs to minimize the chance of false-positive screening hits.

CELL BIOLOGY AND TISSUE ENGINEERING

Wang Y, Waters J, Leung M L, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, Multani A, Zhang H, Zhao R, Michor F, Meric-Bernstam F, Navin N E. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512;2014:155–160.

Breast cancers may be classified into different subtypes according to their cellular phenotype, and this classification correlates with expression of estrogen and progesterone receptors and human epidermal growth factor receptor 2. However, solid tumors commonly harbor genomic heterogeneity among cells comprising a given tumor mass. The present article reports the results of whole-genome sequencing of thousands of individual cells within tumors of different subtypes of breast cancer. To help with genome amplification before sequencing, cells at the G2/M stage of the cell cycle are selected for sequence analysis. These cells have duplicated their DNA in S phase but have not yet divided. They therefore contain twice the diploid DNA amount. Nuclei from cells at G2/M are selected by flow sorting and deposited into wells. The nuclei are then incubated with Φ 29 polymerase to perform multiple displacement amplification. Amplification time is limited to minimize false differences introduced by amplification errors. The amplified DNA is incubated with Tn5 transposase, which fragments the DNA and ligates adaptors for sequencing. The resulting libraries are multiplexed for exome capture or used directly for high-throughput sequencing. The sequence data from single cells are used to compare the kinds of clonal diversity in the different subtypes of breast cancer. Patterns of clonal evolution distinguishing the different breast cancer subtypes are revealed, and different subtypes are shown to have differing mutation rates. Similar methodology for single-cell sequencing is expected to

contribute substantially to knowledge of the pathogenesis of human malignancies.

Ma H, Morey R, O'Neil R C, He Y, Daughtry B, Schultz M D, Hariharan M, Nery J R, Castanon R, Sabatini K, Thiagarajan R D, Tachibana M, Kang E, Tippner-Hedges R, Ahmed R, Gutierrez N M, Van Dyken C, Polat A, Sugawara A, Sparman M, Gokhale S, Amato P, P. Wolf D, Ecker J R, Laurent L C, Mitalipov S. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* 511;2014:177–183.

There are two somatic cell-reprogramming methods for inducing a pluripotent state that is similar to embryonic stem (ES) cells. In the first method, differentiated adult cells are incubated with a transcription factor cocktail to yield induced pluripotent stem (iPS) cells. In the second method, the nucleus of a differentiated adult cell is transferred into the cytoplasm of a fertilized egg, from which the original nucleus has been removed, forming a so-called nuclear transfer (NT) ES cell. Ma et al. report the results of detailed molecular comparisons between in vitro-fertilized ES cells and cells produced by these two methods. The same human subject donated in vitro-fertilized eggs and adult differentiated cells for these studies to maximize genetic similarity of all three cell types for the comparison. Both the NT ES and iPS cells were found to contain comparable numbers of de novo copy-number variants as in vitro-fertilized ES cells, and 30% of the iPS cells had no large duplications or deletions, indicating that iPS cell reprogramming does not cause such genetic defects. DNA methylation analysis and transcriptome profiling, however, showed that iPS cells retain characteristics of their parental adult-differentiated progenitors, whereas NT ES cells are closely similar to in vitro-fertilized ES cells. This indicates that NT produces cells that more faithfully emulate the natural embryonic pluripotent state. This study establishes metrics by which the quality of methods for producing pluripotent cells for analytical and therapeutic purposes may be judged.

Vrtačnik P, Kos Š, Bustin S A, Marc J, Ostanek B. Influence of trypsinization and alternative procedures for cell preparation before RNA extraction on RNA integrity. *Analytical Biochemistry* 463;2014:38–44.

The preparation of high-quality RNA for sequencing or other analytical work is made technically challenging by the nearly ubiquitous presence of ribonucleases capable of degrading the RNA. The authors of the present report observed that RNA could not be prepared from adherent cultured cells if the cells are suspended by treatment with trypsin, a commonly practiced procedure. The reason is found to be that trypsin, purified from animal pancreatic tissue, is contaminated by ribonuclease. The authors suggest circumventing this problem by using trypsin from a nonanimal source or by lysing cells with a phenol-guanidine-based reagent for direct RNA extraction, while the cells are still attached to the culture vessel.

dine-based reagent for direct RNA extraction, while the cells are still attached to the culture vessel.

Bwambok D K, Christodouleas D C, Morin S A, Lange H, Phillips S T, Whitesides G M. Adaptive use of bubble wrap for storing liquid samples and performing analytical assays. *Analytical Chemistry* 86;2014:7478–7485.

Readers of *JBT* typically benefit from well-resourced laboratories that make ample use of disposable containers. However, much research in molecular biotechnology, especially in the area of public health, is conducted in poorly resourced locations, where such containers are inconvenient or unaffordable. Here, Bwambok et al. show that the gas-filled compartments in plastic bubble wrap may serve as containers for a variety of purposes. The bubbles in bubble wrap are transparent, sterile, gas-permeable containers that can serve for storing liquid samples or performing analytical assays. Bubble wrap bubbles may be used as cuvettes for performing absorbance and fluorescence measurements, as culture vessels for growing bacteria and other microorganisms, and as electrochemical cells after inserting carbon electrodes into them. Liquids can be inserted with a syringe needle or pipette tip, after which, the bubbles may be resealed by painting over the hole with nail hardener. These applications are illustrated by culturing *Escherichia coli* and *Caenorhabditis elegans* and by measuring glucose and hemoglobin spectrophotometrically and ferrocyanide electrochemically. This innovation may enable otherwise impracticable research in resource-limited contexts.

DRUG SCREENING AND CHARACTERIZATION

Bachovchin D A, Koblan L W, Wu W, Liu Y, Li Y, Zhao P, Woznica I, Shu Y, Lai J H, Poplawski S E, Kiritsy C P, Healey S E, DiMare M, Sanford D G, Munford R S, Bachovchin W W, Golub T R. A high-throughput, multiplexed assay for superfamily-wide profiling of enzyme activity. *Nature Chemical Biology* 10;2014:656–663.

During screening to identify pharmaceutical enzyme inhibitors, initial trials are usually aimed at optimizing compound potency. Selectivity testing is relegated to later stages of the process and often includes only a restricted panel of optimized compounds and a relatively small number of protein targets. If it were possible to conduct screening of initial compound libraries against a large panel of related proteins, then useful compounds of intermediate affinity but high specificity might be missed less frequently, and less time might be spent on lead compounds whose poor specificity eventually precludes their pharmaceutical use. A screening strategy for this purpose is described by Bachovchin et al. They adopt a competitive, activity-based protein profiling approach, in which purified en-

zymes are coupled to Luminex polystyrene microspheres with a different color for each enzyme. Multiplexed, protein-coupled beads are incubated with a test compound (one compound/well in a screening plate, each compound at multiple different concentrations) and then treated with a biotinylated, activity-based probe and a streptavidin-R-phycoerythrin conjugate. The mixtures are scanned with a Luminex flow cytometer, in which one laser detects the bead color representing the enzyme identity, and another detects the R-phycoerythrin signal representing binding to the test compound. This methodology is applied to assessing inhibitor po-

tency and specificity of serine hydrolase inhibitors. Serine hydrolases constitute an extensive enzyme superfamily with ~240 members in humans. Numerous previously unrecognized, off-target interactions are detected. A highly selective boronic acid-based inhibitor of dipeptidyl peptidase 4 is identified, and the screen is shown to predict successfully the efficacy and safety of this compound in animal models. It is hoped that this methodology will provide a generally applicable, high-throughput approach to screening and optimization for specificity that can be incorporated into the earliest stages of the drug-development process.